Available online at www.sciencedirect.com
SCIENCE DIRECT®
Sensors and Actuators B 96 (2003) 268–282
SENSORS AND ACTUATORS B CHEMICAL
www.elsevier.com/locate/snb

Estimation of chemical and physical characteristics of analyte vapors through analysis of the response data of arrays of polymer-carbon black composite vapor detectors

Brian C. Sisk, Nathan S. Lewis*

Noyes Laboratory, 127-72, Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA 91125, USA

Received 8 November 2002; accepted 9 June 2003

Abstract

Analysis of the signals produced by a collection of organic polymer-carbon black composite vapor detectors has been performed to assess the ability to estimate various chemical and physical properties of analyte vapors based on information contained in the response patterns of the detector array. A diverse array of composite chemiresistive vapor detectors was exposed to a series of 75 test analytes that had been selected from among five different chemical classes: alcohols, halogenated hydrocarbons, aromatics, unsubstituted hydrocarbons, and esters. The algorithmic task of interest was to use the resulting array of response data to assign one of the five chemical class labels to a test analyte, despite having left that analyte out of the model used to generate the class labels. Algorithms evaluated for this purpose included principal components analysis (PCA) and *k*-nearest neighbor (*k*-NN) analysis employing either Euclidean or Mahalanobis distance calculations. Each data cluster that was produced by replicate exposures to an individual analyte was well resolved from all of the other 74 analyte clusters. Furthermore, with the exception of the halide cluster, the analyte response clusters could be robustly grouped into supersets such that each of the five individual chemical classes was well-separated from every other class of analytes in principal component space. Accordingly, using either of the *k*-nearest neighbor algorithms, in excess of 85% of the non-halide test analyte exposures were correctly assigned to their chemical classes, and halides were only routinely confused with aromatics or esters but not with alcohols or hydrocarbons. The detector array response data also was found to contain semi-quantitative information regarding physicochemical properties of the members of the test analyte series, such as the degree of unsaturation of the carbon chain, the dipole moment, the molecular weight, the number of halogen atoms, and type of aromatic ring in the test analytes. The performance in these types of tasks is relevant for applications of a semi-selective array of vapor detectors in situations when no prior knowledge of the analyte identity is available and when there is no assurance that the test analyte will have been contained in the training set database produced by a compiling a library of responses from the detector array.
© 2003 Elsevier B.V. All rights reserved.

Keywords: Sensor arrays; Conducting polymer vapor sensors; Solvent properties

1. Introduction

Arrays of broadly cross-reactive sorption-based detectors have received much recent attention. Typically the sorption detectors are either conducting polymers [1–3] or conductive polymer composites [4–7], polymers that have been impregnated with dyes whose absorption or luminescence signals are sensitive to their environments [8–12], polymer films that have been coated onto surface or bulk resonating crystals [13–15], or polymers that have been coated onto the ends of micromachined cantilevers [16]. In any of these architectures, an analyte elicits a response from many detectors, and in turn each detector responds to many analytes. Pattern recognition algorithms are then used to classify, and in some cases quantify, the analyte of interest [17,18]. Arrays of 5–20 different polymeric sorption detectors have been shown to provide excellent analyte classification and quantification characteristics in a variety of laboratory-based situations [19–22].

Detectors of particular interest in our laboratory are composites that consist of regions of an electrical conductor and regions of an insulating organic polymer [5,23]. Swelling of the polymer by sorption of an analyte induces a reversible, characteristic change in the dc electrical resistance of the detector film. Arrays of such detectors have been shown to provide excellent pairwise resolution between both closely related and diverse analytes, easily resolving between pairs of homologous alkanes, homologous alcohols, H_2O versus

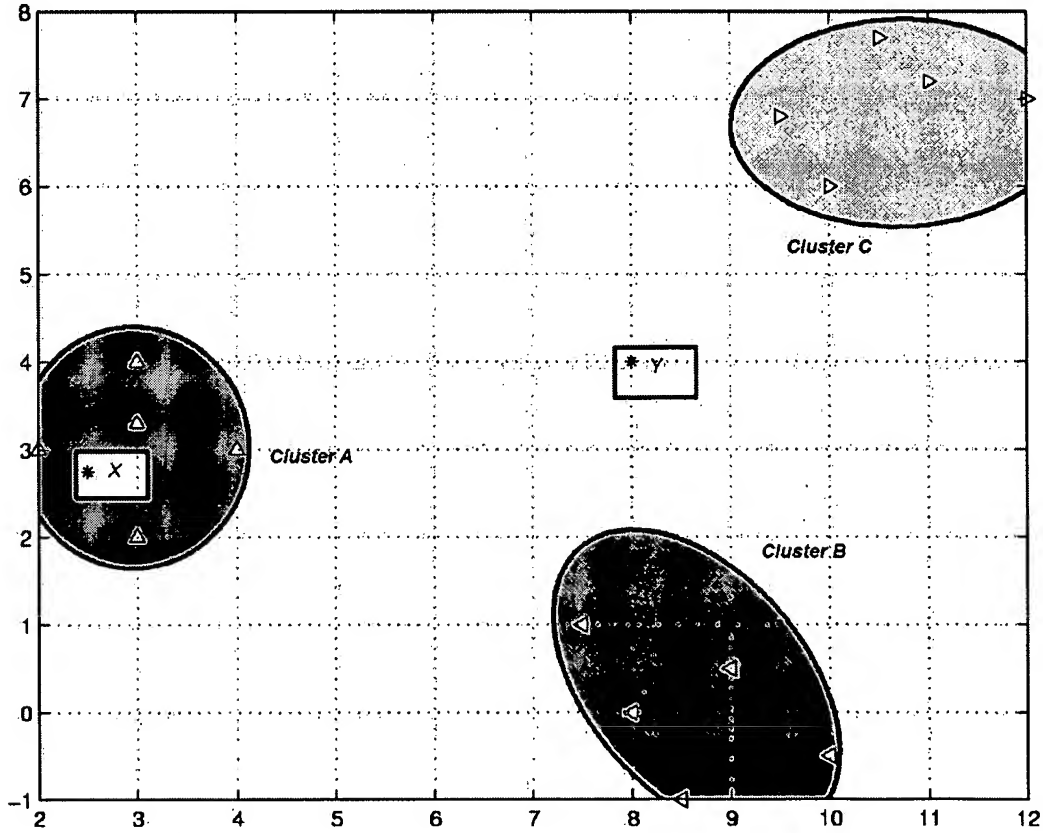
* Corresponding author. Tel.: +1-626-395-6335; fax: +1-626-395-8867.
E-mail address: nslewis@caltech.edu (N.S. Lewis).

0925-4005/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0925-4005(03)00543-4

D_2O, or very similar binary analyte mixtures [6,20,22,24]. Thus, over the timescale of these laboratory experiments, once the detector system has been trained towards these particular odorants, it can readily identify, with a high probability of correct identification and a low rate of false positives, the identity of one of these vapors presented in a subsequent test exposure to the array.

In this work, we have focused on a different question than matching a response pattern to one of the patterns that is already known to be contained in the stored response database for the array of interest. In the present work, we assume that the analyte response information is not in the database, and wish to evaluate what can be deduced about the test analyte through analysis of its array response signals. In many instances, for example, it would be sufficient to be able to classify the general characteristics of an unknown analyte in terms of a chemical classification as an aromatic, aliphatic, chlorinated hydrocarbon, alcohol, ester, or other designated chemical class grouping based on the presence of certain functional groups. Additionally, within such descriptions of analyte classes, it would be of interest to obtain an estimate of selected physicochemical properties of an analyte, such as the value of an analyte's dipole moment, vapor pressure, and/or its molecular volume. Some of this information, such as functional group analysis, is routinely available through analysis of the infrared spectrum of an organic vapor. Other information however, such as molecular volume or substrate binding affinity [25], is more likely to be probed directly by a sorption-based detector than by the amplitude and position of a molecular electromagnetic absorption or emission signal.

As shown in prior work [1,4,6,7], on a sufficiently compositionally diverse array of vapor detectors, each single-component pure analyte will yield its own characteristic response cluster in odor space. The carbon black-polymer composite detectors generally exhibit a response that is linear with analyte concentration, so that the response cluster for a given analyte is maintained for normalized data over a wide range of analyte concentrations [6]. Each response cluster can furthermore typically be differentiated with relatively high resolution from the response clusters produced by exposure of the detector array to all of the other analytes in the training set. The question of interest is whether a decision surface can be drawn in the resulting n-dimensional odor space (where n equals the number of detectors used) such that if a test analyte exposure falls inside of the decision surface, the test analyte can also be



Scheme 1. Description of clustering as implemented in this study. Three distinct analyte clusters (A, B, and C), consisting of five members each, are shown in a simulated principal components analysis plot. Each of the three clusters is completely separated from the others. Also included are two "unknown" test analytes, X, and Y. X is completely contained by the boundary that defines class A, and each of the three points nearest to X is a member of class A. Therefore, X is assumed to belong to class A. Y is well outside the boundaries of all three classes, and the analytes nearest to it do not correspond to the same analyte class. Therefore, Y is likely not a member of any of classes A, B, or C.

Table 1
Analytes presented to the detector array

Alcohols	Halides	Aromatics	Hydrocarbons	Esters
Methanol	1-Chlorobenzene	Benzene	Cyclooctane	Isopropyl acetate
Cyclopentanol	1-Bromobutane	Propyl benzene	*n*-Hexane	Butyl acetate
2-Butanol	Cyclohexyl chloride	*m*-Xylene	*n*-Octane	Pentyl acetate
1-Pentanol	1,1,2-Trichloroethane	*o*-Xylene	*n*-Decane	Methyl acetate
2-Pentanol	1-Bromopentane	*p*-Xylene	3,3-Dimethyl 1-butene	Isobutyl acetate
3-Pentanol	3-Chloro 2-methyl propene	Isopropyl benzene	*n*-Heptane	*trans*-2-Hexenyl acetate
Isopropanol	1-Chloropropane	Ethyl benzene	*n*-Nonane	Hexyl acetate
Ethanol	2-Chlorobutane	Toluene	Cyclopentane	Isopentyl acetate
1-Butanol	1-Fluorobenzene	1,2,4-Trimethyl benzene	2,2,4-Trimethyl pentane	Ethyl propionate
2-Methyl 1-propanol	1-Iodopropane	2,6-Lutidine	Cyclohexane	Propyl acetate
3-Methyl 1-butanol	2-Bromo 2-methylpropane	2-Picoline	*n*-Pentane	*sec*-Butyl acetate
2-Methyl 2-butanol	1-Iodobutane	Pyridine	2,5-Dimethyl 2,4-hexadiene	Isopentyl propionate
2-Propen-1-ol	Chloroform	Anisole	2-Methyl 2-butene	Pentyl butyrate
1-Hexanol	Methylene chloride		7-Methyl 1,6 octadiene	Isopentyl benzoate
2-Methyl 3-buten-2-ol	1-Chlorobutane		1,7-Octadiene	Ethyl butyrate
			Cyclopentene	
			Cyclooctene	

correctly identified with high probability as being a member of the same chemical class as the training set of analytes that is contained inside the decision surface. An example for a hypothetical two-dimensional odor space is shown in Scheme 1, in which analyte X is correctly assigned to one of three possible analyte clusters, whereas analyte Y is not successfully assigned to a class. Additional class boundaries can in principle be formulated to describe other physicochemical properties of the analytes of interest, such as molecular volume, dipole moment, etc.

To evaluate these possibilities, arrays of organic polymer-carbon black composite chemically sensitive resistors were exposed to a series of single-component organic vapors. The vapors were members of one of five distinct chemical classes, as indicated in Table 1. In a "leave-one-out" approach [17], subsets of these analytes formed the models and databases that were used in conjunction with data analysis algorithms to extract physicochemical information on the test analyte. Principal components analysis (PCA) and *k*-nearest neighbor (*k*-NN) analysis using both Euclidean and Mahalanobis distance calculations were evaluated for their ability to assign correctly the chemical class, degree of unsaturation, number of halogen atoms, nature of the aromatic ring, dipole moment, and molecular volume of each of the 75 test analyte vapors.

2. Experimental

The detector array consisted of two copies each of 20 compositionally distinct polymer-carbon black composite chemically sensitive resistors (Table 2), for a total of 40 detectors. Detector films were cast from mixtures of 80% polymer and 20% by weight of carbon black (Black Pearls 2000, Cabot Inc.), as described previously [5]. The detector films were deposited between two Au leads that had been evaporated onto a glass slide, and the array was housed in a stainless steel assembly that was connected by Teflon tubing to a computer-controlled, calibrated vapor generation and delivery system.

The set of 75 pure single-component analyte vapors was formed from approximately 15 members from each of five distinct analyte classes: alcohols, halides, aromatics, hydrocarbons, and esters (Table 1). Due to a limited number of sol-

Table 2
Polymers used to fabricate the polymer-carbon black composite detector array

1	Poly(ethylene oxide)
2	Poly(ethylene oxide)-*co*-poly(amidoamine), diblock gen. 4[a]
3	Poly(ethylene-*co*-vinyl acetate) (45% vinyl acetate)
4	Poly(ethylene oxide)-*co*-poly(amidoamine), diblock gen. 1[b]
5	Poly(styrene-*b*-butadiene)
6	Kraton G[c]
7	Polyvinyl carbazole
8	Kraton D[c]
9	Poly(vinyl acetate)
10	Poly(diphenoxyphosphazene)
11	Polycaprolactone
12	Polychloroprene
13	Polysulfone
14	Polyaniline-0.5-HDBSA[d]
15	Poly(vinyl pyrrolidone)
16	bis(cyanoallyl polysiloxane)
17	Poly(4-vinyl phenol)
18	Poly(styrene-*co*-allyl alcohol)
19	Poly(methyloctadecyl siloxane)
20	Ethyl hydroxyethyl cellulose

[a] PEO-PAMAM diblock copolymer, with 5000 M_W linear PEO and generation 4.0 PAMAM dendrimer (total M_W = 8420).

[b] PEO-PAMAM diblock copolymer, with 5000 M_W linear PEO and generation 1.0 PAMAM dendrimer (total M_W = 5230).

[c] Commercial polymer from Shell Corp.

[d] Polyaniline with a 0.5 fraction of all amine sites protonated by hexadecyl benzene sulfonic acid (HDBSA).

vent bubblers available in the experimental apparatus, data collection was divided into many runs, with each run consisting of exposures to 8 out of the 75 analytes. The first run consisted of 10 exposures of the detector array to each of the first 8 analytes (from top to bottom as tested) from Table 1, with the analytes presented in random order. The next run was made up of two randomly selected analytes from run 1, as well as six new analytes, Table 1. The selection process was then repeated again, and a new set of eight analytes was created from a pair of analytes from the previous run and six new analytes. The selection process was continued until each analyte had been included at least once in a run. Then, another round of runs was performed with each run containing two randomly chosen analytes from four of the five analyte classes, and this process was repeated until each analyte had been included in a second run. In this way, each analyte was presented to the detector array on at least two different occasions. Analytes were all presented at a fixed (0.020) fraction of their vapor pressure at room temperature, $21 \pm 1\,^{\circ}C$, to insure a constant vapor phase analyte activity throughout the runs.

Each analyte exposure consisted of a 3 min pre-exposure period to allow measurement of a stable baseline resistance, followed by a 5 min period of analyte flow during which the steady-state differential resistance change of the detector was recorded. A 7 min post-exposure period allowed the detector resistances to return to baseline after each analyte exposure. Resistance data were recorded using a multiplexing Keithley multimeter and a data acquisition computer as described previously [20].

Baseline correction of the data was performed by fitting a regression line to the first 10 points of the pre-exposure resistance readings, and correcting all subsequent data points by the difference in the value of the regression fit at the time of the measurement of that data point and at $t = 0$. A single descriptor, the relative differential resistance change, $\Delta R/R_b$, was used in the analysis of the response of each detector to an analyte exposure. The resistance values upon analyte exposure were measured as the average over 10 data points after reaching equilibrium. ΔR was measured as the difference in resistance between the equilibrium response states before (R_b) and during analyte exposure. Each analyte exposure therefore produced a 40-dimensional vector, as follows:

$$X = \sum_{i=1}^{40} c_i x_i \tag{1}$$

With X_i being the $\Delta R/R_b$ value of the Xth detector. Prior to quantitative data analysis, principal components analysis was used to visualize portions of the unnormalized, 40-dimensional detector array response data. The first two principal components contained 66% of the total variance of the 40-dimensional data while 76% of the total variance was contained in the first three principal components; thus, the data were visualized using the first three principal components as axes. The three-dimensional data clouds were rotated manually while the data were viewed along a fixed axis to assess the separation between clusters for various tasks of interest.

The k-nearest neighbor approach was used to obtain a quantitative measure of analyte classification ability in different tasks. First, the mean response vectors for each of the 75 analytes were calculated by averaging the unnormalized array responses recorded during the replicate exposures to each analyte. A leave-one-out approach was then used, and a model data set was formed from the mean response vectors produced by exposures to 74 of the 75 total analytes. No response data from the analyte of interest was included in the construction of this model database of response vectors. For each individual exposure to the test analyte of interest, up to seven of the nearest mean response vectors in the model database were then identified. The procedure was repeated for each of the 20 exposure data points for the analyte of interest. Finally, the entire process of model database construction, excluding data for the analyte of interest, and assessing distances to other mean analyte response vectors in the database, was repeated for each of the 75 analytes studied in this work.

Distance measurements in the k-NN analysis were made using both Euclidean and Mahalanobis distances. The Euclidean values were calculated simply by determining the distances between two points in the 40 dimensional space, with no prior scaling or normalizing of the $\Delta R/R_b$ response data values:

$$r^2 = (x - \mu)^T (x - \mu) \tag{2}$$

Here, r is the Euclidean distance between an unknown analyte response vector x and the mean vector response μ, and the superscript T indicates the transpose operation. Mahalanobis distances differ from Euclidean distances in that they are calculated on data for which each of the 40 individual descriptors is first autoscaled across the entire data set. When autoscaling, each of the 40 dimensions is mean-centered, then divided by the variance of that particular dimension [17]:

$$r^2 = (x - \mu)^T \Sigma^{-1} (x - \mu) \tag{3}$$

Here, r indicates the Mahalanobis distance between x and μ, and Σ^{-1} represents the inverse of the covariance matrix derived from the original $m \times n$ data matrix derived from m analyte exposures to an array of n detectors. The Mahalanobis distance approach has the advantage in principle that noisier dimensions (detectors) do not unduly dominate the distance calculated between two data points in the 40-dimensional response space. However, it is possible for this data transformation to introduce artifacts in certain cases, so classification was evaluated using both this approach and a conventional Euclidean distance measurement. The class of each test data point was assigned to the class that was represented by the majority of the nearest k mean response vectors, with k varying as 1, 3, 5 or 7. Each

analyte was assigned to only be a member of a single-class (Table 1). No class assignment was made to a data point in instances when the nearest neighbor data points selected did not produce a majority class consensus.

Calculations of the molecular volume and dipole moments of various analytes in the gas phase were performed using the Cerius[2] software package. Volumes were determined using a probe of 1 Å, and the calculated charges were Mullikan charges. Structures were determined by performing a series of steepest descent energy minimizations and Mullikan charge calculations, which ultimately resulted in convergence of both energy and charge. Calculations of the dipole moment and molecular volume of each of the test analytes were performed on these minimized structures.

3. Results

3.1. Class assignment

Fig. 1 presents a plot of the analyte response data for all 75 analytes as projected onto the first three principal components of the 40-dimensional detector $\Delta R_{max}/R_b$ response space. For clarity, only the mean response vector termini, obtained from the average of multiple exposures to each analyte, are displayed. Each of the 75 different analytes produced a response cluster that was clearly separable from the response cluster of every other analyte investigated in this work. As is apparent from Fig. 1, the mean vector response termini clustered into four well-defined, mutually separated regions. Three of these regions individually contained the mean vector response termini for the alcohols, hydrocarbons, and esters, respectively. In contrast, substantial overlap was present between the region that contained vector response termini produced by exposure of the detector array to halides and the regions containing either aromatic or ester organic vapors.

The class assignment performance enabled by the detector array data was quantified using k-NN analysis (Table 3). Analytes that are members of two analyte classes (fluorobenzene, chlorobenzene, and isopentyl benzoate) were excluded from this analysis. Despite a high degree of overlap between the clusters produced by the halides and both the aromatics and esters, the mean correct class assignment probability for the 72 single-class analytes tested was 0.76–0.82 using Euclidean distances and was 0.76–0.81 using Mahalanobis distances. Classification rates were largely insensitive to whether $k = 3$, 5, or 7 nearest neighbors was used. The data in Table 3 suggested that using seven neighbors instead of three decreased the correct classification rate more than the incorrect classification rate, and significantly increased the number of non-classified exposures.

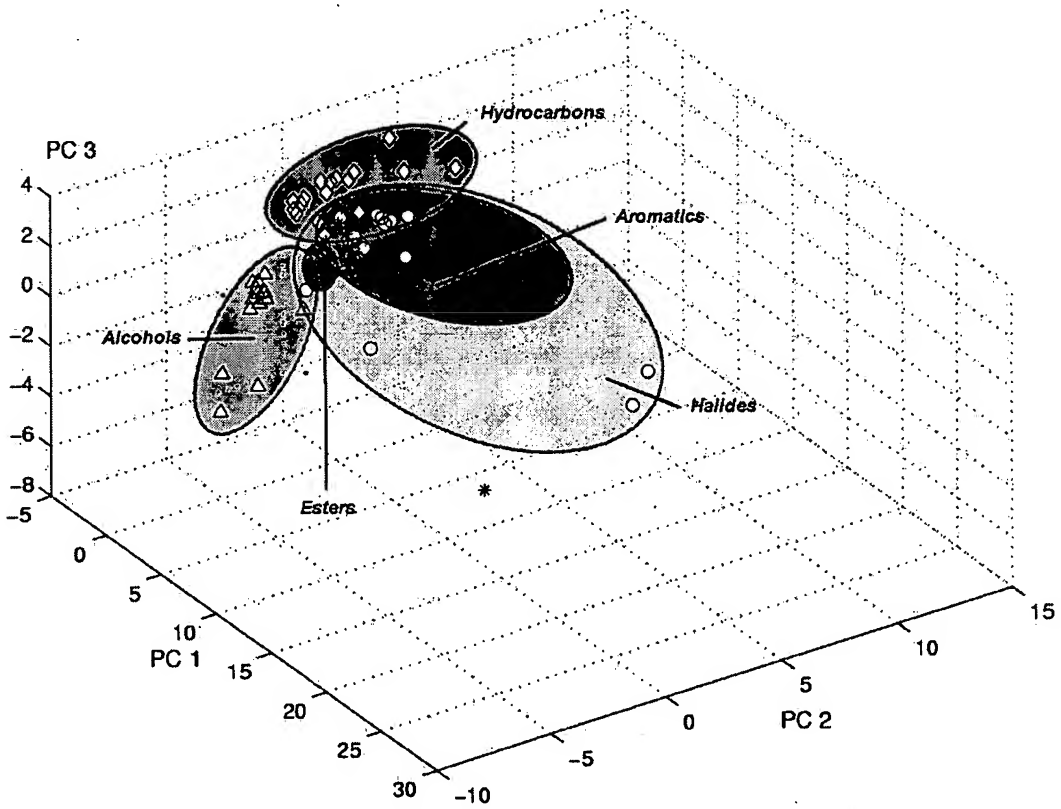


Fig. 1. Principal components analysis plot of the mean response vector termini for each analyte used in the study, each of which corresponds to one of five analyte classes: alcohols, halides, aromatics, hydrocarbons, and esters. Approximate classification boundaries have been drawn around each of the five classes.

Table 3
Fractions of analyte exposures correctly classified using *k*-nearest neighbor analysis

Neighbors		Mahalanobis	Euclidean
1	Correct	0.7773	0.7893
	Incorrect	0.2227	0.2107
	Non-classified	0.0000	0.0000
3	Correct	0.7860	0.7713
	Incorrect	0.1613	0.1767
	Non-classified	0.0527	0.0520
5	Correct	0.7753	0.7547
	Incorrect	0.1553	0.1493
	Non-classified	0.0693	0.0960
7	Correct	0.7540	0.7400
	Incorrect	0.1373	0.1500
	Non-classified	0.1087	0.1100

Classifications using three neighbors were preferable to those derived from a single neighbor, however, as the increase in the non-classification rate resulted almost exclusively from a decrease in the incorrect classification rate. Therefore, the three-nearest neighbor algorithm was used in all further analysis.

Table 4 report the performance by analyte class in the form of confusion matrices for the class assignments using Mahalanobis distances and three nearest neighbors, where each (X, Y) cell in the table indicates what fraction of exposures belonging to an analyte class X was assigned to the analyte class Y. Perfect classification performance would produce the identity matrix in this representation of the success of class prediction from the detector array response data. Table 4a shows the results of all five classes, while Table 4b displays the results having left the halides out of the analysis. The overall the correct classification rate for all 72 single-class analytes using three nearest neighbors and Mahalanobis distances was 0.80; excluding halides (retaining all alcohols, aromatics, hydrocarbons, and esters), this rate increased to 0.88. The analytes with the worst classification performances among the alcohols, aromatics, hydrocarbons, and esters corresponded to isopropyl benzene, cyclopentene, 2,5-dimethyl 2,4-hexadiene, methyl acetate, and trans-2-hexenyl acetate, which resulted in error rates of 1.0, 0.80, 0.80, 1.0, and 0.7, respectively. Removing these five analytes (as well as the halides) from the set resulted in an overall correct classification rate (Mahalanobis distances, three neighbors) of 0.95. Thus, by removing 23 of the 75 analytes, the combined rates of non-classification and error was cut by three-fourths.

3.2. Determination of chemical information in addition to class identity

The structure of the data displayed in Fig. 1 suggests that classification of additional analyte properties should be possible, because the majority of the non-halide frequently misclassified analytes were polyfunctional, unsaturated, or had low molecular volumes relative to most members of their respective classes. Examples are provided by the responses of isopropyl benzene, cyclopentene (small, cyclic, unsaturated), 2,5-dimethyl 2,4-hexadiene (polyunsaturated), methyl acetate (small), and trans-2-hexenyl acetate (the only ester tested with an unsaturated hydrocarbon chain), as labeled in Fig. 2. Consequently, further analysis was performed to determine whether such analyte-specific information could be isolated in a systematic fashion from the array vector response data.

Fig. 3 displays a principal components analysis plot of 20 exposures each of the detector array to *n*-hexane, *n*-heptane, *n*-octane, 2,2,4-trimethyl pentane, 2-methyl 2-butene, 2,5-dimethyl 2,4-hexadiene, 1,7-octadiene, and 7-methyl 1,6-octadiene. These analytes contain various degrees of unsaturation in the hydrocarbons yet minimize differences in their molecular weights. The broad, diffuse cloud of array response data in Fig. 3 that was produced by the unsaturated hydrocarbons was clearly separable from the tighter cluster of response data that was produced by the saturated hydrocarbons. The only overlap between the

Table 4
Confusion matrices developed from *k*-nearest neighbor analysis using three neighbors and Mahalanobis Distances

	Alcohols	Halides	Aromatics	Hydrocarbons	Esters
(a)					
Alcohols	0.9733	0	0	0	0.0267
Halides	0.0077	0.4577	0.1654	0.0192	0.2000
Aromatics	0	0.0885	0.8154	0.0077	0.0346
Hydrocarbons	0.0147	0.0265	0.0324	0.8441	0.0235
Esters	0.0500	0.0714	0.0036	0	0.8464
(b)					
Alcohols	0.9767		0	0	0.0233
Aromatics	0		0.8538	0.0192	0.0808
Hydrocarbons	0.0147		0.0588	0.8382	0.0500
Esters	0.0821		0.0143	0.0036	0.8714

The value of any cell X, Y indicates what fraction of exposures from the analyte in row X was classified as the analyte in column Y. Note that fluorobenzene, chlorobenzene, and isopentyl benzoate have been omitted from this analysis as they are members of two classes.

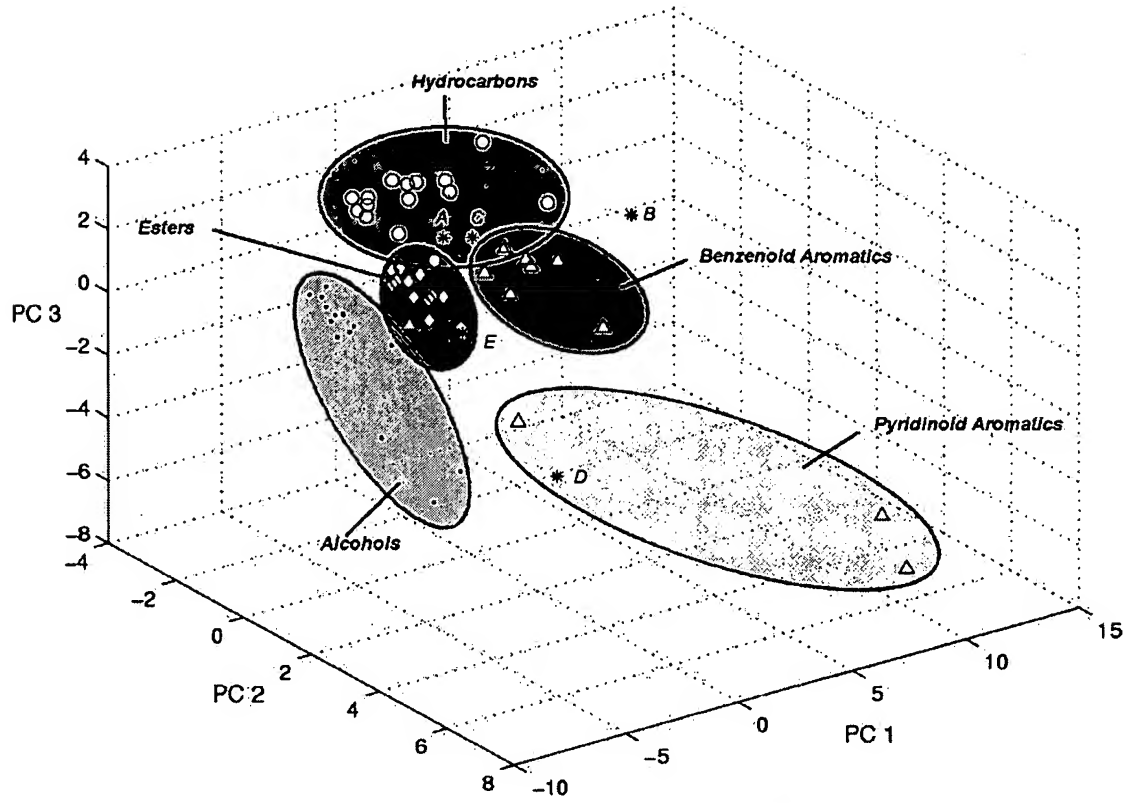


Fig. 2. Principal components analysis plot of the mean vector termini for each alcohol, aromatic, hydrocarbon, and ester except those that were members of multiple classes (chlorobenzene, fluorobenzene, and isopentyl benzoate). The five analytes that were most frequently misclassified by *k*-nearest neighbor analysis using Mahalanobis distances and three nearest neighbors, isopropyl benzene (A), cyclopentene (B), 2,5-dimethyl 2,4-hexadiene (C), methyl acetate (D), and trans-2-hexenyl acetate (E), are specifically labeled in the plot as (✱).

two clusters (from the vector angle shown) arose from half of the exposures to 7-methyl 1,6-octadiene.

A confusion matrix analysis of the *k*-NN determined (Mahalanobis, three nearest neighbors) classification performance in this task indicated that the fraction of saturated hydrocarbons correctly identified as such was 0.96; unsaturated hydrocarbons were correctly classified with a probability of success of 0.79. Hydrocarbons which produced responses that were located outside of a single, tight cluster were predicted, with a high degree of success, to be unsaturated. In fact, 10 of the 17 mistakes made in classification of the unsaturated hydrocarbons corresponded to 7-methyl 1,6-octadiene, the only unsaturated hydrocarbon that overlapped with the cluster of saturated hydrocarbons.

Fig. 4 displays a principal components analysis plot of 20 exposures to non-aromatic analytes that contained single or multiple halide groups. The multi-functional halides employed were chloroform, dichloromethane, and 1,1,2-trichloroethane; additionally, all non-aromatic halides (but excluding fluorobenzene and chlorobenzene) from Table 1 were included in the analysis. With the exception of four outlier data points that arose from some of the exposures to 1-iodopropane, the analytes that contained single halide functionality were well-separated from analytes that had multiple halide functional groups. This separation persisted despite a wide disparity in molecular weights (78.5–119 g mole^{-1}) and molecular volumes (81.4–126 Å^3) within the group of analytes that contained a single halide functionality. Additionally, although the multi-functional halides had no more than two carbons, a good deal of overlap existed between their molecular weight and molecular volume ranges and those of the monofunctional analytes (96.9–133.5 g mole^{-1} and 58.6–92.6 Å^3, respectively).

Fig. 5 displays a principal components analysis plot of the clustering that arose from the mean vector response termini from each analyte in the halide and aromatic vapor sets. The clustering is divided into four categories: benzenoid aromatics, pyridinoid aromatics, mono-functional halides, and multi-functional halides. The pyridinoid analytes produced signals that were well-separated from the signals produced by the benzenoid analytes. The clustering of the benzenoid aromatics from the pyridinoids persisted despite the presence of anisole (methoxybenzene), which is similarly polar, so a simple polarity argument is not sufficient to explain the clustering. Furthermore, adding chlorobenzene and fluorobenzene to the set resulted in their responses falling well within the benzenoid aromatic/monofunctional halide cluster. The slight separation between the halide and aromatic clusters (Fig. 1) was largely produced by the presence of the multi-functional halides and pyridinoid aromatics. The

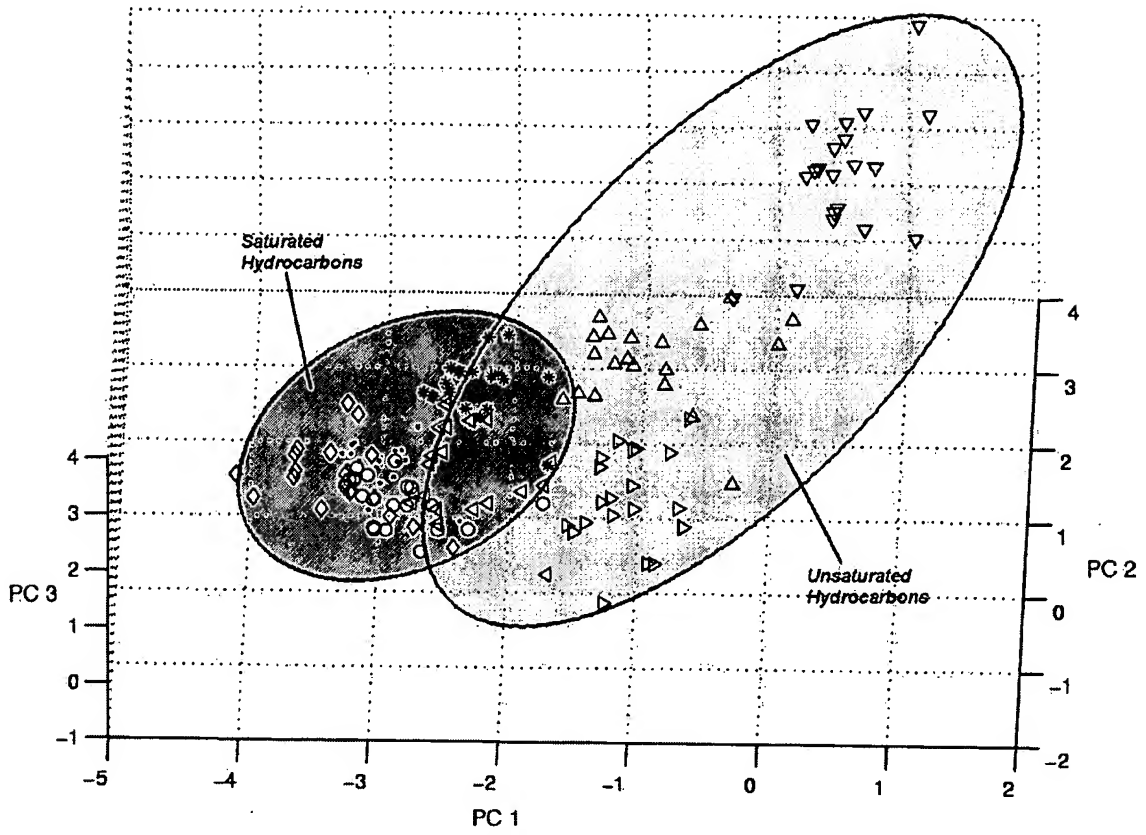


Fig. 3. Principal components analysis plot of data collected from eight hydrocarbons: *n*-hexane, *n*-heptane, *n*-octane, 2,2,4-trimethyl pentane, 2,5-dimethyl 2,4-hexadiene, 2-methyl 2-butene, 7-methyl 1,6-octadiene, and 1,7-octadiene. All 20 exposures for each analyte are shown. The first four analytes, which are saturated, are represented in the plot by (●), (✱), (◇), and (○), respectively. The last four, which are unsaturated, are represented by (△), (▽), (◁), and (▷), respectively. Sufficient separation is achieved between the two clusters to determine saturation with a success rate of over 80%.

observation that the pyridinoid aromatics were easily separated from the benzenoid is not necessarily surprising, given the increased basicity of the pyridinoid structure, but it is interesting from a classification point of view that the two groups are reasonably well-separated from each other. Furthermore, it is surprising that the degree of overlap between the benzenoids and single-functional halides is so high, even more so than indicated by the PCA and *k*-NN results for aromatics relative to halides.

3.3. Determination of physical characteristics

Although the class properties dominated the clustering of the mean response vector termini of the analytes investigated, physical properties such as the molecular weight of the analyte also contributed to the form of the data in principal components space. Fig. 6 shows a principal components analysis plot of the molecular weight of the analyte, in which the various molecules have been binned by size into five categories. The first four categories, or bins, had upper limits of 92.654, 109.05, 122.47, and 143.65 Å^3, respectively, and the final bin had no upper limit. As in Fig. 1, mean vector response termini were used for clarity. The bin cutoffs were chosen to ensure equal populations of 150 exposures in all five bins. The smallest molecules were the most likely to be found far from the central data cluster, while the largest molecules produced mean vector response termini that were close to, or included in, the central cluster regardless of the functionality contained in the analyte of interest.

Fig. 7 displays a plot of the average molecular volume of the analytes in each bin as a function of the average Mahalanobis distance between the mean response vector terminus for each analyte in the bin and the overall centroid of the total data set. For this plot, 10 bins were used, with the first nine bins having upper limits of 80.957, 92.654, 100.06, 109.05, 113.30, 122.47, 132.86, 143.65, and 162.85 Å^3, respectively. The nearly monotonically decreasing curve formed by the data in Fig. 7 indicates the good correlation between the average Mahalanobis distance and the average molecular volume of the analytes in the respective bins.

While there is no real clustering in Fig. 6, the general trend of distance versus molar volume persists nonetheless, and can be useful at least as negative identification. For analytes in the bin with the smallest volume in the 10-bin set, averaging 63.8 Å^3, no single exposure had a distance from the full-data centroid that was within the first quartile of the distance values for the entire dataset. Similarly, for analytes in the each of the four largest bins, averaging 127–179 Å^3,

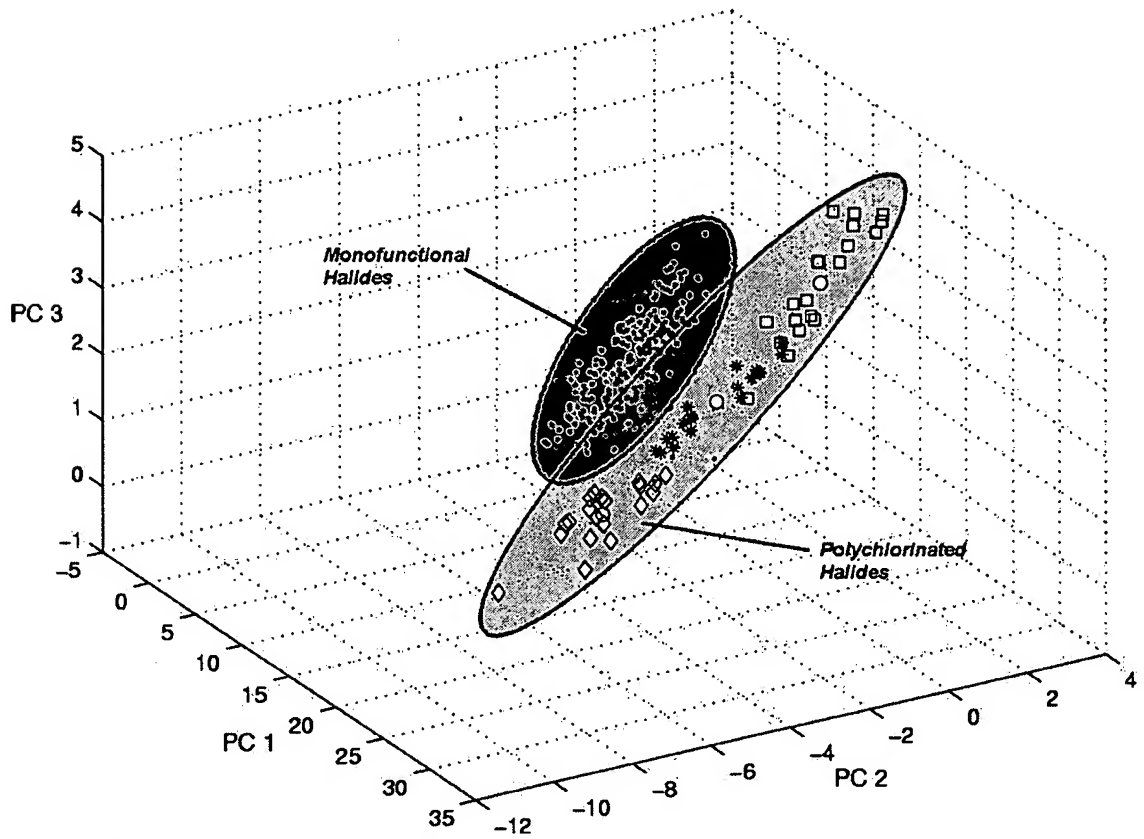


Fig. 4. Principal components analysis plot of data collected from all of the halides used in this study (20 exposures each). The polychlorinated analytes, including 1,1,2-trichloroethane, dichloromethane, and chloroform are represented in the plot by (□), (✱), and (◇), respectively. Mean vector response termini for the polychlorinated clusters are denoted by (○), with A representing 1,1,2-trichloroethane, B representing CH_2Cl_2, and C representing $CHCl_3$. All of the other halides are represented by (●). The polychlorinated analytes are well-separated from the main group of halides.

the fraction of exposures having a distance above the third quartile was below 0.133. Therefore, analytes very near the centroid are unlikely to be small, and analytes far from the centroid are not likely to be large. Distances that fall outside of the middle half of all exposures should thus add useful information to the task of classifying analyte exposures. Table 5 shows a statistical summary of the binned principal components data, with mean volumes and distances from the full-data centroid reported, as well as probabilities of a single exposure from a given bin being within the first quartile, median, and third quartile distances of the full-data centroid.

The approximate dipole moment of the analyte could also be extracted from the array response data. The results are displayed in Fig. 8, for which the data from the standard principal components analysis plot were binned by separating those analytes with dipole moments <1.0 D from those analytes having dipole moments >1.0 D. Tighter clustering was observed between the low-dipole moment analytes, largely

Table 5
Statistical data for principal components analysis data grouped by volume

Bin #	Average volume (Å^3)	Average distance from centroid	Probability above first quartile distance	Probability above median distance	Probability above third quartile distance
1	63.77	170.8	1.00	0.987	0.72
2	87.26	44.27	0.973	0.747	0.573
3	95.86	57.56	0.787	0.613	0.347
4	104.5	19.86	0.773	0.613	0.320
5	112.3	19.15	0.787	0.680	0.133
6	117.1	38.59	0.600	0.427	0.227
7	126.6	11.83	0.573	0.280	0.0133
8	138.7	11.74	0.600	0.213	0.0400
9	152.1	10.76	0.520	0.227	0.000
10	179.4	14.89	0.893	0.213	0.133

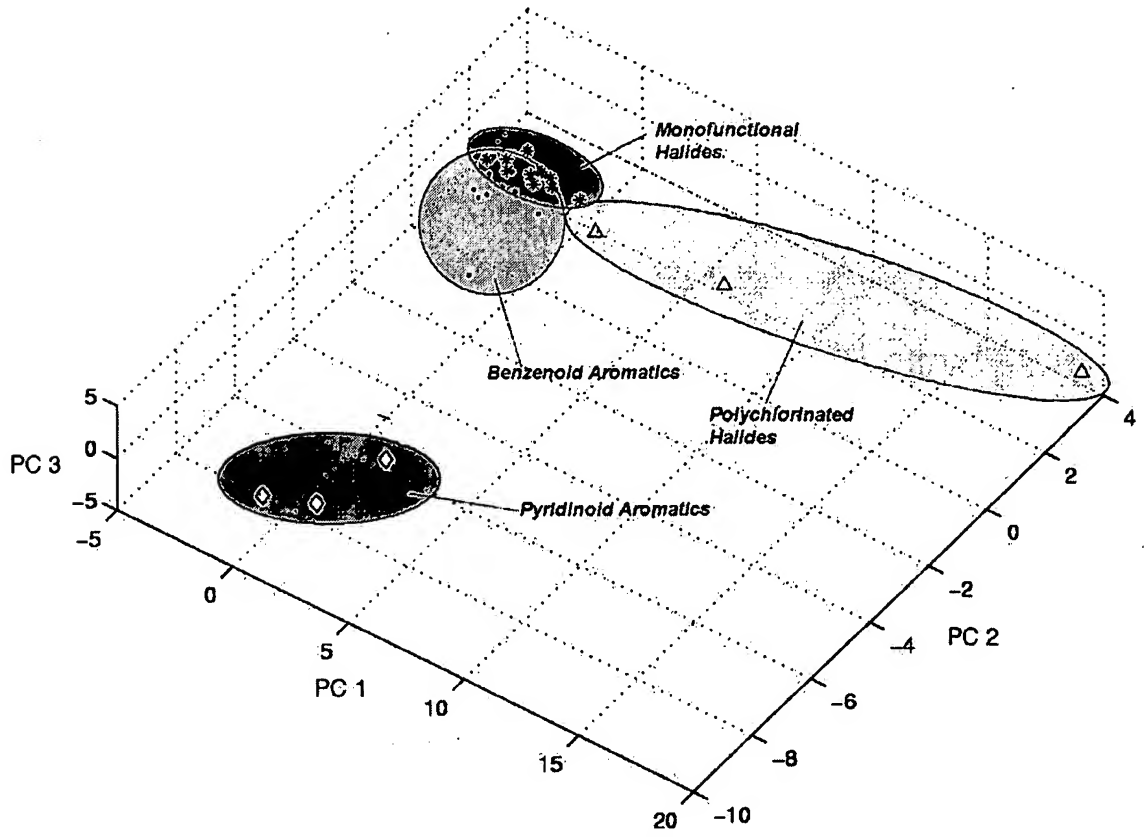


Fig. 5. Principal components analysis plot of the mean vector response termini for the halides and aromatic analytes used in this study, excluding chlorobenzene and fluorobenzene. Pyridinoid aromatics (◇) are well-separated from benzenoid aromatics (●), as are polyfunctional halides (△) from monofunctional halides (✱). The bulk of the separation between the halides and aromatics in this study arises from the presence of polyfunctional halides and pyridinoid aromatics—separation between benzenoid aromatics and monofunctional halides is poor.

consisting of hydrocarbons and non-heteroatom aromatics, than between the other analytes. Of course, the dipole moment is correlated with the analyte class in this particular test set. However, Fig. 8 indicates that the benzenoid aromatics are much more similar to the hydrocarbons, viewed from the shown vector, than they are to the pyridinoid aromatics. Also, a great deal more clustering was observed among the non-polar analytes, between classes, than was displayed among the polar analytes. Because of the trends regarding analyte class and molecular volume that have already been extracted to a reasonable degree of accuracy, it is impossible to achieve perfect clustering with regard to dipole moment. However, again, it appears that the analytes that have the highest distances from the central cluster have dipoles over 1.0 D. Therefore, it should be possible to choose a distance value above which the analytes primarily have dipoles above 1.0 D.

For this analysis, the "enrichment" of high-dipole analytes above a certain distance cutoff from the centroid is displayed in Fig. 9 as a function of how many exposures meet the distance cutoff value. The enrichment is calculated as $\varepsilon = (A - E)/(A + E)$, where A is the ratio of the number of exposures of high-dipole analytes to low-dipole analytes above the chosen cutoff, and E is a similar ratio for the general population. For example, 164 of the 750 exposures have distances over 26.5 (arbitrary units) from the full-data centroid. Of these, 135 have dipoles above 1.0 D and 29 have dipoles below 1.0 D, yielding a ratio of $135/29 = 4.66$ for the 164 exposures with the furthest distances from the centroid. Because 490 of the exposures in the entire dataset have exposures above 1.0 D, with 260 below, the overall dataset has a ratio of 1.88. Therefore, the 164 furthest exposures have a high-dipole "enrichment" of $(4.66 - 1.88)/(4.66 + 1.88) = 0.425$. Enrichment values range from $\varepsilon = 0$ to 1, with 1 representing complete separation of the high- and low-dipole analytes at a given cutoff value. Fig. 9 shows that the enrichment of high-dipole analyte exposures increases with larger distance cutoff values and consequently fewer exposures reaching the cutoff.

Analysis of the analytes both having dipoles below 1.0 D and distances far from the full-data centroid is naturally limited to aromatics and hydrocarbons. The low-dipole moment analytes with the five highest distances were pentane, cyclopentane, 3,3-dimethyl 2-butane, benzene, and cyclopentene. Of these nonpolar outliers, none has more than six carbons. Of the 12 polar analytes furthest from the centroid, included were naturally the pyridinoids (basic) and the multi-functional halides (multiple heteroatoms). Other

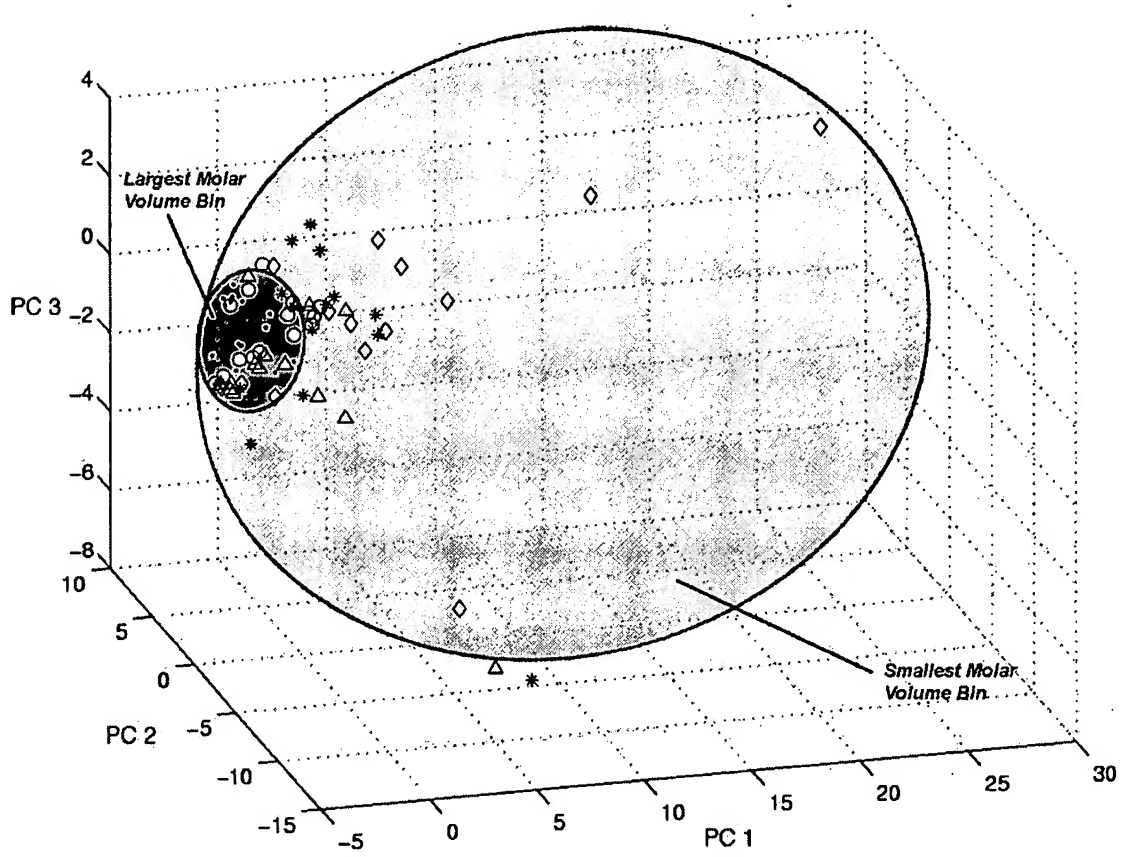


Fig. 6. Principal components analysis plot of the mean vector response termini for all analytes included in this study. The analytes were binned by molecular volume, with 150 analyte exposures in each of five bins. The largest analytes are included in the bin denoted by (●), with size decreasing with the bins marked by (○), (△), (✱), and finally (◇). Solvent envelopes which include all analytes in the bin have been included in the plot for the smallest and largest bins. The envelope that surrounds all of the analytes in the bin with the smallest volumes is much larger than that which surrounds the analytes in the bin with the largest volumes. This shows that the small analytes in this study tend to be more varied in their properties than the larger analytes.

than these, which comprised 6 of the 12, were methanol, ethanol, allyl alcohol, 1-chloropropane, methyl acetate, and isopentyl benzoate—the smallest three alcohols, the smallest halide, the smallest ester, and the only aromatic ester. As such, the trends with regard to dipole and molecular volume relative to distance from the full-data centroid must be considered together—the analytes furthest from the centroid will likely be small, highly polar, or both.

4. Discussion

To our knowledge, little information is available to date in a leave-one-out study protocol regarding whether mapping into functional groups and/or geometric descriptors of a molecule can be robustly performed from the response patterns produced by an array of semi-selective sorption-based vapor detectors. At sufficiently high analyte concentration, the responses of a variety of detectors are sufficiently distinct that unique identification is possible for most analytes. However, because the responses of semi-selective detectors by nature depend on a large number of factors, including molecular volume, branching, dipole, hydrogen bonding, aromaticity as well as many others, the ability to extract any one of these parameters to the exclusion of the others, or at least by limiting them, has not been fully elucidated to date.

Within a single analyte class, isolating certain variables (such as size or saturation) proved successful because there were, in general, few differences between the analytes in the set except the variable of interest. For instance, within a homologous series of alcohols, only molecular weight and branching of the chain separated the 12 unsaturated members of the set, with three also possessing a unit of unsaturation. Monofunctional halides were successfully separated from bi- or tri-functional halides, benzenoid aromatics were distinct from pyridinoids, and saturated hydrocarbons were not readily confused with unsaturated hydrocarbons. This is largely due to the fact that, of the many factors to which a semi-selective detector is responsive, frequently only one varies significantly within a particular chemical identification task, so the chemical sub-classes (such as saturated hydrocarbons) are largely homogeneous relative to the differences between the sub-classes.

However, successfully classifying a set of analytes that differ not only in analyte class but also with regard to many other parameters was found to be more difficult. Among the

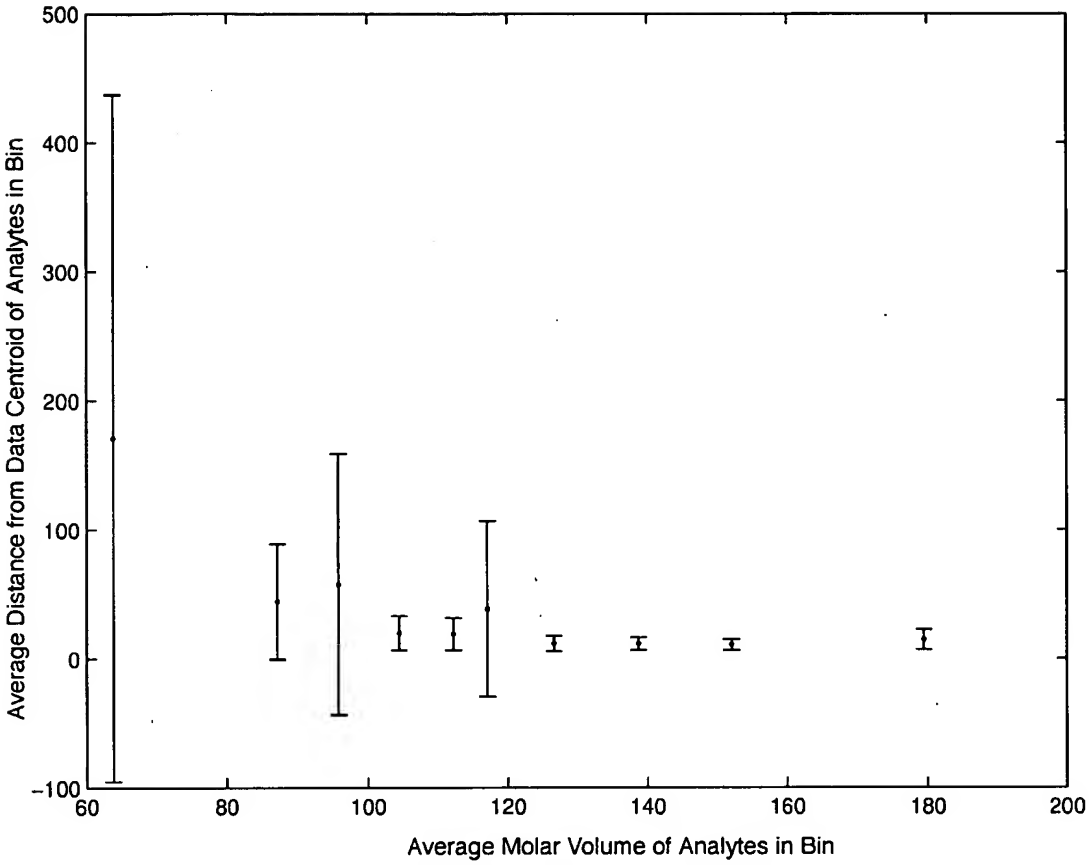


Fig. 7. Average distance from the centroid of the entire dataset vs. average molecular volume. For this plot, 10 bins were used, each consisting of 75 analyte exposures. The largest average distance was found with the bin having the smallest average molecular volume, and the four bins with the highest average molecular volumes had the lowest average distances.

non-halide classes, the bulk of the mistakes in *k*-NN analysis was contained in five analytes, such as methyl acetate or 2,5-dimethyl 2,4-hexadiene, that differed significantly from their base analyte classes. Because classification proved successful for a high proportion of analyte exposures, unique class characteristics such as aromaticity, hydrogen bonding, and others found in this analyte set must largely overwhelm considerations such as saturation, branching, and volume differences, which vary considerably within classes without leading to significant rates of misclassification.

The comparative difficulty in correctly identifying analyte class members that differ from their respective classes is not surprising. However, the result does not imply that recognizing analytes within diverse classes is impossible, as found from the high rate of correct classification among the pyridinoid aromatics, which differ greatly from the benzenoid. Rather, a sufficiently robust analyte basis set must be established to ensure that a sufficiently similar neighbor exists in the analyte database. Expecting correct classification of pyridine as aromatic, for example, would be unreasonable if lutidine or picoline were not present in the set. However, with a well-developed analyte basis set, a great deal of within-class diversity can be tolerated before poor classification performance is obtained.

Though it is beneficial that the detector set employed in this analysis proved more sensitive to solvent class characteristics than physical characteristics, it is not clear that this trend can be generalized to all detector sets. It is certainly conceivable that the overwhelming variable could have been molecular volume. Had this been the case, efficient clustering along class boundaries would likely have not been possible. An array of detectors comprised of a homologous series of polymers that differed only in chain length, such as poly(1-alkenes), might prove more sensitive to molecular volume than to chemical characteristics. Because the detector array in this study was designed to ensure that a variety of polymers with diverse characteristics was represented, it should not be surprising, then, that the detector array used in this study is most sensitive to chemical class.

The heightened sensitivity of the detector set toward low molecular weight analytes, measured by their propensity to deviate from the full-data centroid, can be explained by the differences between the high- and low-volume analytes within an analyte series. As most of the analytes are monofunctional, increasing volume within a homologous series typically results from an increase in the number of carbon atoms in an analyte molecule. It should not be surprising, then, that two analyte class clusters are nearest

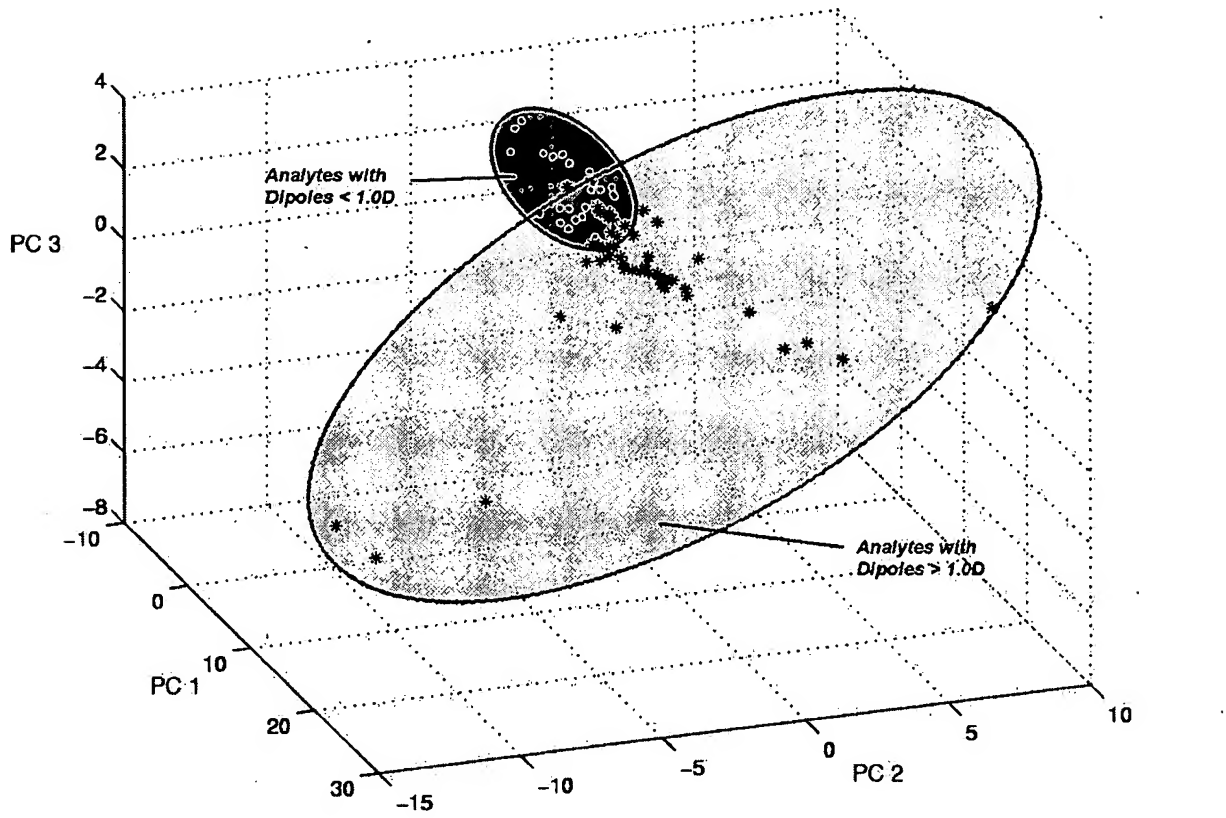


Fig. 8. Principal components analysis plot of the mean vector response termini for each of the analytes used in this study. For this plot, the analytes were binned by dipole moment, with (●) representing the analytes with dipole moments <1.0 D, and (✱) representing the analytes with dipole moments >1.0 D. The low-dipole analytes were well-clustered and separated from the high-dipole analytes. The envelope surrounding the low-dipole analytes is significantly smaller than that surrounding the high-dipole analytes. This result suggests that the low-dipole analytes tested are less diverse than high-dipole analytes as measured by the detectors used in this study.

each other where their members are the largest, and that the smallest members should radiate out from a common center. Considering a single example, *n*-chlorohexane and *n*-hexanol are expected to be more similar to each other than are chloromethane and methanol, simply because *n*-chlorohexane and hexanol have more hydrocarbon character in common than do methanol and chloromethane. Therefore, the clustering of large analytes near the center of the analyte space can be explained without considering any particular bias within the detector array.

A similar argument based on increasing hydrophobic character within homologous series cannot explain the clustering of low-dipole analytes, however, because none of the halides, esters, or alcohols studied had a dipole less than 1.0 D, regardless of chain length. Instead, it is more likely that the low-dipole molecules lack diversity because they cannot form two of the three principal analyte–polymer interactions, namely hydrogen bonding and dipole–dipole. Additionally, heteroatoms increase the diversity of analytes, and the low-dipole groups, hydrocarbons and benzenoid aromatics, are completely devoid of heteroatoms, since heteroatoms tend to raise an analyte's dipole over 1.0 D. Therefore, most low-dipole analytes are forced to consist almost solely of carbon and hydrogen, and there is certainly a limit to how much diversity can be achieved among relatively small molecules consisting of only these two elements. Those analytes that possess both heteroatoms and low-dipoles tend have those heteroatoms aligned in a highly symmetric fashion, and none of the analytes with heteroatoms in this study met that requirement. An example of an analyte that meets this criterion is triazine, which contains three nitrogen atoms in a D_{3h} configuration.

While heteroatom deficiency may be sufficient to explain the clustering of low-dipole analytes, it is possible that this effect is partially due to the detector set chosen. Because extremely hydrophobic detectors are frequently difficult to dissolve and spray-cast into viable detectors, there are few detectors in the array based on polymers with particularly low-dipoles. If the detector array is disproportionately composed of polymers having higher dipoles, it may not be surprising that there is a greater diversity of analyte responses among the more polar analytes.

The polymer-carbon black composite vapor detectors used in this study have previously been shown to produce a response that is linear with analyte concentration. Hence, the analysis performed herein is essentially independent of

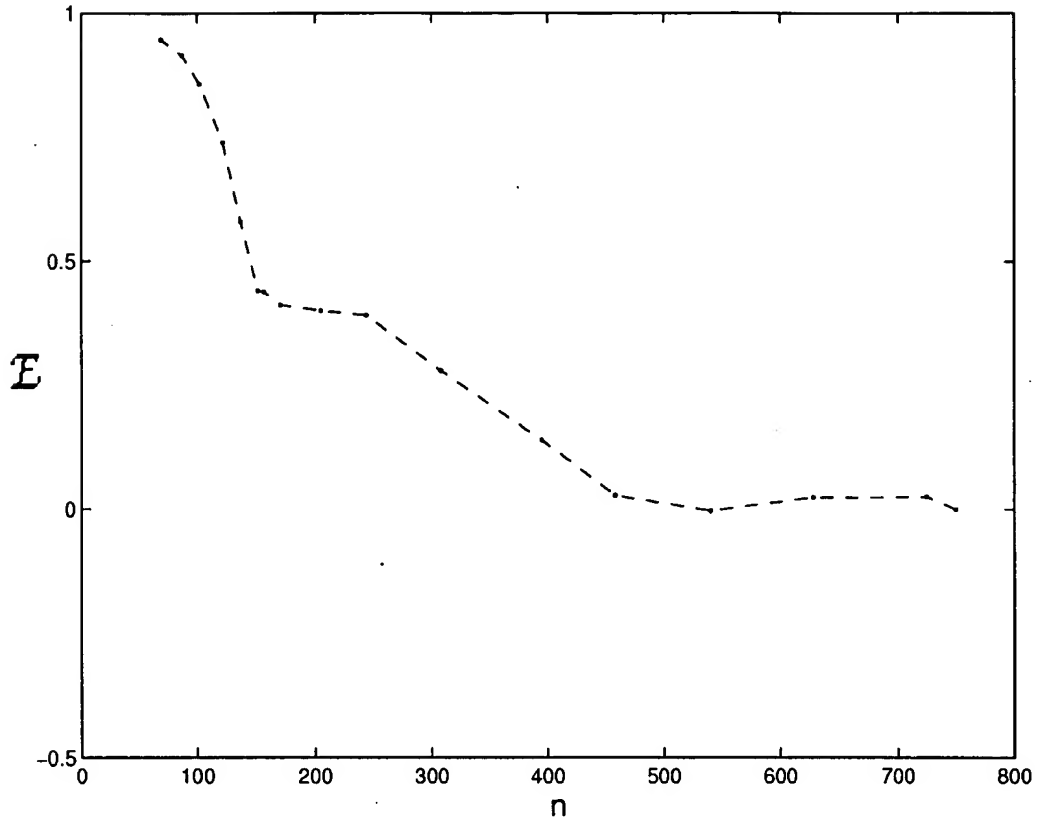


Fig. 9. Enrichment, E, of high-dipole exposures (>1.0 D) in an analyte exposure set in which each exposure is above a certain distance cutoff from the full-data centroid vs. the number of exposures, n (out of 750 total) that meet that cutoff. An enrichment of $E = 0$ corresponds to the dipole distribution similar to that found in the full-data set; an enrichment of $\mathcal{E} = 1$ corresponds to a dataset consisting only of high-dipole analytes. As can be seen, higher (more selective) distance cutoffs result in datasets consisting almost solely of high-dipole moment analyte exposures.

the concentration of analyte. [6]. Development of a robust model for non-linearly responding detectors is expected to be less straightforward, and likely requires collection of larger datasets along with modeling the detector response as a function of analyte concentration [26]. Such complications are minimized through the use of carbon black-polymer composite detectors, which facilitates a straightforward development of classification models for various chemical and physical properties of the analytes of interest.

5. Conclusions

In addition to identifying with a high degree of confidence analytes to which it has been previously exposed, an array of semi-selective polymer-carbon black composite detectors is also capable of qualitatively describing analytes to which it has never been previously exposed. k-NN and principal components analysis indicate that, with the exception of the halide analyte class, the detector array used in this work is capable of assigning with a high degree of confidence class descriptors to analytes. In addition to the overall analyte class, a variety of chemical sub-class information can also be extracted from the raw data, including hydrocarbon saturation, mono-functionality versus poly-functionality among halides, and the nature of aromatic rings (benzenoid versus pyridinoid). Additionally, in certain cases information regarding the size and dipole moment of molecules can be determined. Ultimately, it is possible that in many cases sufficient knowledge of an unknown analyte's characteristics can be established through the extraction of chemical and physical parameters to allow tentative identification of analytes that have not been previously encountered by the detector array.

Acknowledgements

We acknowledge the NIH, NSF, and an Army MURI for their generous support of this work.

References

[1] J.W. Gardner, A. Pike, N.F. Derooij, M. Koudelkahep, P.A. Clerc, A. Hierlemann, W. Gopel, Integrated array sensor for detecting organic solvents, Sens. Actuators B, Chem. 26 (1995) 135–139.

[2] P.N. Bartlett, P.B.M. Archer, S.K. Ling-Chung, Conducting polymer gas sensors. 1. Fabrication and characterization, Sens. Actuators B, Chem. 19 (1989) 125–140.

[3] H.V. Shurmer, J.W. Gardner, P. Corcorran, An electronic nose-sensitive and discriminating substitute for a mammilian olfactory system, Sens. Actuators B, Chem. 1 (1990) 256–260.
[4] M.S. Freund, N.S. Lewis, A chemically diverse conducting polymer-based electronic nose, Proc. Natl. Acad. Sci. U.S.A. 92 (1995) 2652–2656.
[5] M.C. Lonergan, E.J. Severin, B.J. Doleman, S.A. Beaber, R.H. Grubbs, N.S. Lewis, Array-based vapor sensing using chemically sensitive, carbon black-polymer resistors, Chem. Mater. 8 (1996) 2298–2312.
[6] E.J. Severin, B.J. Doleman, N.S. Lewis, An investigation of the concentration dependence and response to analyte mixtures of carbon black/insulating organic polymer composite vapor detectors, Anal. Chem. 72 (2000) 658–668.
[7] D. Hodgins, The development of an electronic nose for industrial and environmental applications, Sens. Actuators B, Chem. 27 (1995) 255–258.
[8] T.A. Dickinson, J. White, J.S. Kauer, D.R. Walt, A chemical-detecting system based on a cross-reactive optical sensor array, Nature 382 (1996) 697–700.
[9] J. White, J.S. Kauer, T.A. Dickinson, D.R. Walt, Rapid analyte recognition in a device based on optical sensors and the olfactory system, Anal. Chem. 68 (1996) 2191–2202.
[10] T.A. Dickinson, K.L. Michael, J.S. Kauer, D.R. Walt, Convergent, self-encoded bead sensor arrays in the design of an artificial nose, Anal. Chem. 71 (1999) 2192–2198.
[11] N.A. Rakow, K.S. Suslick, A colourimetric sensor array for odour visualization, Nature 406 (2000) 710–713.
[12] C. Ronot, Detection of chemical vapors with a specifically coated optical-fiber sensor, Sens. Actuators B, Chem. 11 (1993) 375–381.
[13] D.S. Ballantine Jr., S.L. Rose, J.W. Grate, H. Wohltjen, Correlation of surface acoustic wave device coating responses with solubility properties and chemical structure using pattern recognition, Anal. Chem. 58 (1986) 3058–3066.
[14] S.L. Rose-Pehrsson, J.W. Grate, D.S. Ballantine, P.C. Jurs, Detection of hazardous vapors including mixtures using pattern recognition analysis of responses from surface acoustic wave devices, Anal. Chem. 60 (1988) 2801–2811.
[15] S.J. Patrash, E.T. Zellers, Characterization of polymeric surface acoustic wave sensor coatings and semiempirical models of sensor responses to organic vapors, Anal. Chem. 65 (1993) 2055–2066.
[16] H.P. Lang, M.K. Baller, R. Berger, C. Gerber, J.W. Gimzewski, F.M. Battiston, P. Fornero, E. Meyer, H.J. Gunderhodt, An artificial nose based on a micromechanical cantilever array, Anal. Chim. Acta 393 (1999) 59–65.
[17] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[18] P.C. Jurs, G.A. Bakken, H.E. McClelland, Computational methods for the analysis of chemical sensor array data from volatile analytes, Chem. Rev. 100 (2000) 2649–2678.
[19] J. Park, W.A. Groves, E.T. Zellers, Vapor recognition with small arrays of polymer-coated microsensors. A comprehensive analysis, Anal. Chem. 71 (1999) 3877–3886.
[20] S.V. Patel, M.W. Jenkins, R.C. Hughes, G. Yelton, A.J. Ricco, Differentiation of chemical components in a binary solvent vapor mixture using carbon/polymer composite-based chemiresistors, Anal. Chem. 72 (2000) 1532–1642.
[21] B.J. Doleman, M.C. Lonergan, E.J. Severin, T.P. Vaid, N.S. Lewis, Quantitative study of the resolving power of arrays of carbon black-polymer composites in various vapor-sensing tasks, Anal. Chem. 70 (1998) 4177–4190.
[22] M.C. Burl, B.C. Sisk, T.P. Vaid, N.S. Lewis, Classification performance of carbon black-polymer composite vapor detector arrays as a function of array size and detector composition, Sens. Actuators B, Chem. 87 (2002) 130–149.
[23] A.J. Matzger, C.E. Lawrence, R.H. Grubbs, N.S. Lewis, Combinatorial approaches to the synthesis of vapor detector arrays for use in an electronic nose, J. Comb. Chem. 2 (2000) 301–304.
[24] T.P. Vaid, M.C. Burl, N.S. Lewis, Comparison of the performance of different discriminant algorithms in analyte discrimination tasks using an array of carbon black-polymer composite vapor detectors, Anal. Chem. 73 (2001) 321–331.
[25] T.P. Vaid, N.S. Lewis, The use of electronic nose sensor responses to predict the inhibition activity of alcohols on the cytochrome p-450 catalyzed p-hydroxylation of aniline, Bioorg. Med. Chem. 8 (2000) 795–805.
[26] G.C. Osbourn, J.W. Bartholemew, A.J. Ricco, G.C. Frye, Visual-empirical region-of-influence pattern recognition applied to chemical microsensor array selection and chemical analysis, Acc. Chem. Res. 31 (1998) 297–305.